

SEARCHING TEXT AND DATA via COMMON GEOGRAPHY

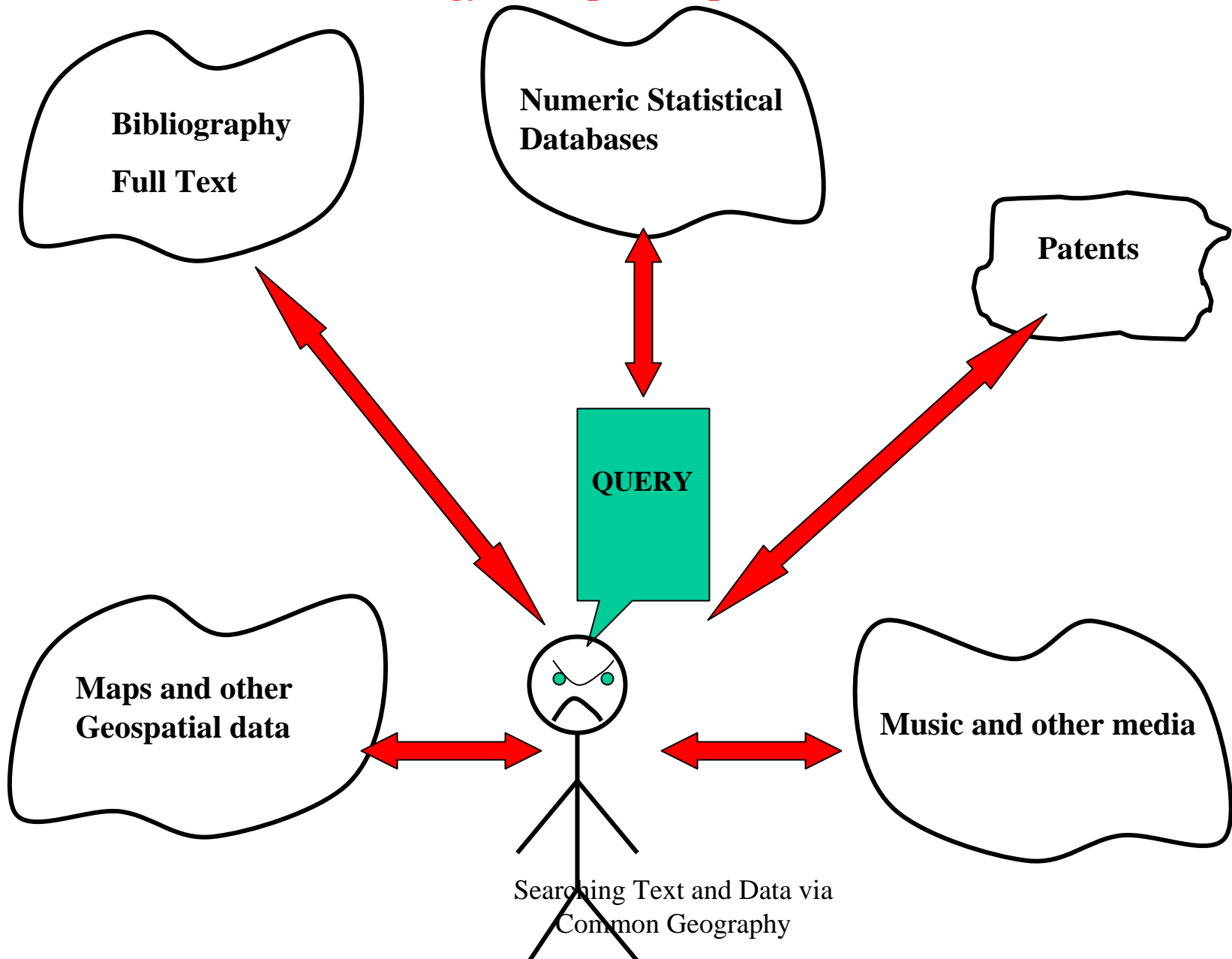
- **Geographic Information Retrieval: Searching Text and Data via Common Geography**
- **IASSIST 2002 Conference, June 12-14, 2002, Storrs, CT**
- **PIs - Fredric C. Gey, Michael Buckland, Aitao Chen, Ray Larson, University of California, Berkeley**
- **Students:**
 - **Vivien Petras, Natalia Perleman**
- **Work performed under Institute for Museum and Library Services (IMLS) *national leadership grant 1999-2002* and**
- **DARPA research contract N66001-97-8541;AO#F477: *Search support for unfamiliar metadata vocabularies (1997-2001)* and**
- **IMLS proposal *Going Places in the Catalog: Improved Geographic Access***

MOTIVATION

- **The purpose of social science research is question answering:**
 - What are the population characteristics of Visalia California?
 - What is the history of Visalia?
- Answering such questions requires cross-genre search
- **Currently only humans can search cross-genre information**
- Search across genres requires **metadata linkage**
- **Geography is a major linkage between numbers which describe a place and text which explains it**
- Gazetteers uniquely identify places in space

HETEROGENEOUS DIGITAL INFORMATION SEARCH

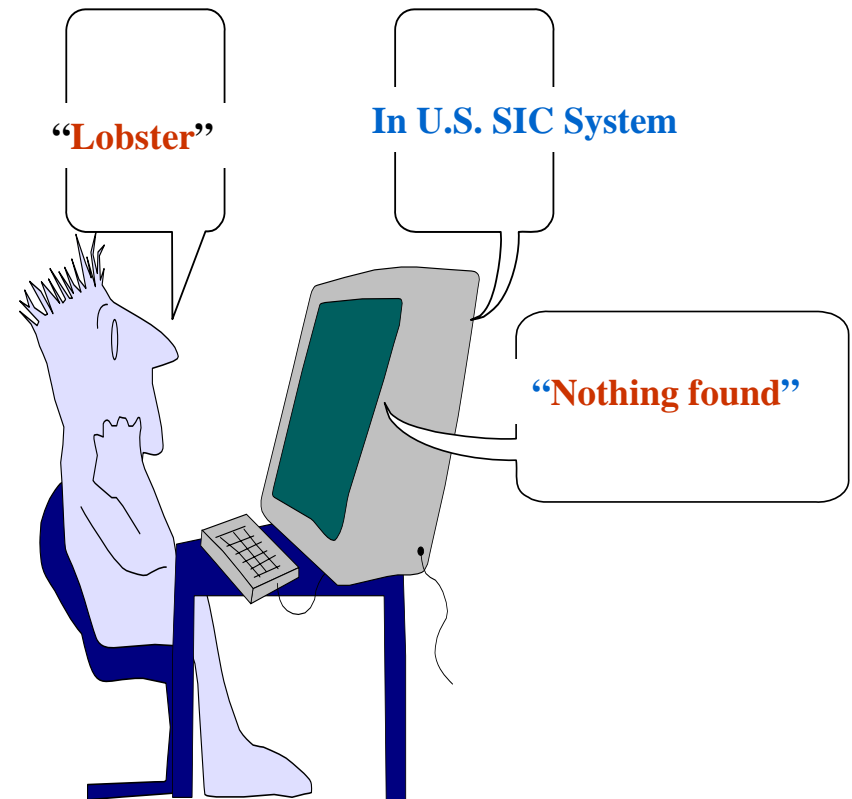
Current Search Technology (multiple independent searches without search aids)



Searching Text and Data via
Common Geography

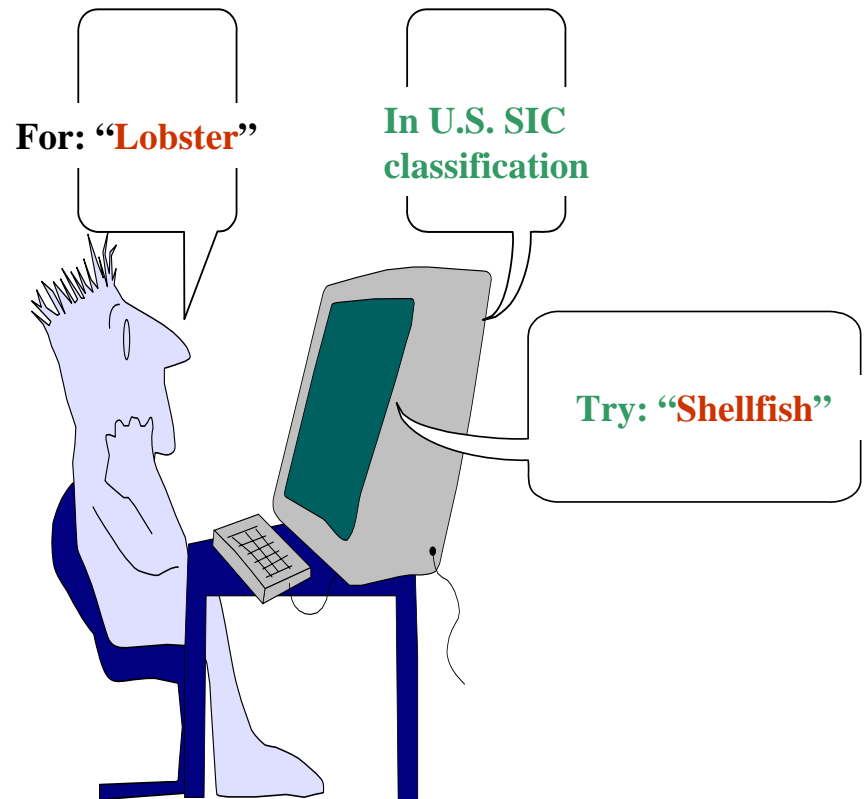
OUR PRIOR RESEARCH LINKED TEXT AND NUMBERS: U.S. STANDARD INDUSTRIAL CLASSIFICATION SYSTEM

- **U.S. Standard Industrial Classification System (SIC)**
- **Used to classify and aggregate industrial activity in the U.S.**
- **Codes defined by Office of Management and Budget**
- **Descriptions are incomplete**



MINING TEXT TO SEARCH NUMBERS WITH ENTRY VOCABULARY TECHNOLOGY

- Mapped between ordinary language and specialized classifications
- Implemented using text categorization techniques
- Required text collections which have been manually indexed
- Preserves and leverages investment in creation of complex classification structures



Kinds of metadata

- **Physical metadata for location of data**
- **Structural metadata to identify the logical structure of the data**
 - **matrix of age(31) by race(5) by sex(2)**
- **Measurement metadata (units of measure, universe of discourse)**
- **Semantic metadata which describes the meaning of the data (usually a descriptive segment of text which identifies the meaning of a classification, e.g. value label)**
- **Locational metadata for identifying the geospatial and temporal aspects of data**
 - **latitude, longitude, altitude, time**

Why Geographic Searching Presents Special Opportunities

- **Place is pivotal for interdisciplinary inquiry:** anthropologists, economists, historians, military strategists, political scientists have common ground in space, place, spatial changes over time.
- Geography links numeric information about a place with textual information which elucidates the place.
- One can do new and unique things in library catalogs:
 - “find books describing the history of all towns within 30 miles of Visalia CA”

LINKING NUMERIC AND TEXT DATABASES

- Effective search requires **clues and evidence**
- Numeric/statistical datasets
 - Limited textual descriptive information
- Library catalogs (bibliographic databases)
 - Ambiguity of places (**Vienna, VA** or *Vienna, Austria*)
- Improve search between these evidence-poor and ambiguous databases by **metadata linkage via common geography**

Ambiguity of Geography in Text

- **Ambiguity of identical names: Alameda (city) or Alameda (county)? Galicia (region of Spain) or Galicia (region of Poland)?**
- **Different transliterations for from non-Roman scripts: Peking is a variant spelling of Beijing.**
- **Different names in different languages: Deutschland, Allemagne, Germany**
- **Name changes: Bombay is now Mumbai; St Petersburg became Leningrad became St Petersburg again.**
- **Political changes disrupt place-name stability: Poland ceased to be a country (late 18th century) when partitioned between Austria-Hungary, Prussia and Russia; Prussia is no longer a country**
- **Footprint: even when a place's name is stable the area it denotes may not be.**

Gazetteer Characteristics can be Exploited

- **Gazetteers map place-names to unique positions: Washington DC**
 - **Latitude: 38.90505 North – Longitude: 77.01616 West**
(geographic centroid of the polygon representing the city)
- **Gazetteers may contain other useful features:**
 - **Feature type: city, lake, church, bridge, ...**
 - **Larger region in which the place resides: county, state, country**
 - **Additional items** such as a **map** showing the place-name
 - **Information about the time-range** in which a particular place was or is current
- **Gazetteers allow spatial relationships between named places to be calculated and utilized**
 - **Numeric data: How many people live within 30 miles of Visalia?**

Our Gazetteer Research Project Proposal

- Make use of library catalog MARC record geographic features in new ways
- Connect numeric data with library record information using gazetteers as intermediate metadata
- Utilize gazetteer information to extend library catalog search
- Display catalog search results in map displays to enable users to visualize search results
- Exploit feature type metadata to develop more complex spatial queries: “What books on travel and description concern places within 25 miles of dams in California?”
- Utilize map interfaces to allow users to generate visual queries
- Extend searches from catalogs to other geo-referenced datasets, e.g. museum and digital cultural heritage collections.

SUMMING UP

- **Semantic metadata for numbers often has limited textual content**
- **Semantic metadata for text may be ambiguous in time or space**
 - **Aberdeen Scotland or Aberdeen Maryland**
 - **St. Petersburg or Leningrad**
 - **Prussia is no longer a county**
- **Gazetteers may be exploited to uniquely specify location and provide clues for disambiguation.**
- **Linking numbers, text and gazetteers allows for new ways to search library catalogs**

Further Information:

- www.sims.berkeley.edu/research/metadata
- www.sims.berkeley.edu/research/seamless
- **Michael Buckland and Ray Larson (buckland,
ray @sims.berkeley.edu)**
- **Fredric Gey (gey@ucdata.berkeley.edu)**

